GME to the Moon, or how r/WallStreetBets squeezed GameStop

Atri Bhattacharyya
EPFL
atri.bhattacharyya@epfl.ch

Aurel R. Mader
EPFL
aurel.mader@epfl.ch

Ludovica Schaerf
EPFL
ludovica.schaerf@epfl.ch

ABSTRACT

In January 2021, the Reddit community r/WallStreetBets was attributed for causing a once-in-a-decade short squeeze on the GameStop stock. In this paper, we analyze the interactions (posts, comments) within the community to determine whether the community could indeed have been responsible, and whether the actions of the community represent a mass movement of individual players.

KEYWORDS

GME, Reddit, Network Analysis, Stock Market, Regression Analysis

ACM Reference Format:
Atri Bhattacharyya, Aurel R. Mader, and Ludovica Schaerf. 2021. GME to the Moon, or how r/WallStreetBets squeezed GameStop. In *Proceedings of ACM Conference (Project Report).* ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

Communities on social media have the power to affect large-scale social movements since social media allows people to connect far beyond their local communities, and with unprecedented speed.

In this paper, we investigate the role of a community on the social media platform Reddit, r/WallStreetBets (r/WSB), in the short squeeze of the publicly traded stocks of GameStop, a retailer of video games with physical stores across the world.

In January 2021, the stock prices of GameStop ($GME stock) saw a bizarre price movement. Whereas the stock had consistently traded for less than $10 over the previous year, the price rose meteorically, peaking at $347 [17]. Being a physical retailer in the age of giant digital game distributors such as Steam and Epic Games, GameStop was widely regarded as a relic of the past and predicted to slowly fade into irrelevance (parallels exist to the decline of the video rental chain Blockbuster, being supplanted by video streaming services such as Netflix).

There were two major factions betting on $GME. On one side, hedge funds had made short contracts amounting to more than 140% of $GME's stocks [1]. On the other side, prominent members of the r/WallStreetBets community encouraged retail investors to buy the stock en masse, believing it to be undervalued [6]. During the ensuing price spike, it is reported that short sellers lost over $6 billion [1]. While long term stock holders from the Reddit community made significant gains, a large number of Reddit users who entered the market near the peak have suffered large losses.

Our interest in the $GME short squeeze revolves around understanding the dynamics of the Reddit community (called a *subreddit* in Reddit terminology). To counter the billions invested by hedge funds, the investments from the subreddit would have to match it. Given the limited means of retail investors, a large number of them would have to pool resources to cause the short squeeze, as often claimed [6, 17]. Or the short squeeze might also have been driven and orchestrated by a few financially strong actors. This introduces our first research question: whether the /r/WallStreetBets community is dominated by a few highly influential individuals, or whether the community is more egalitarian and influenced by a larger number of members.

For our second research question, we ask whether the actors in the subreddit were actually responsible for the short squeeze, and whether the publicly available information from the subreddit can be used to measure the impact of the subreddit on price fluctuations of $GME.

2 LITERATURE REVIEW

The short squeeze of the Gamestop stock has been vastly covered by the media and recent academic research [4, 6, 10, 11, 17, 18]. Prior research analyzes the short squeeze by looking at the interest in the GME stocks as visible on Google trends [11, 18] and pays special attention to the influence of Google trends on the trading volume of the GME stock [18]. The influence of the subreddit r/WallStreetBets in the GME phenomenon has been investigated before by trying to quantify the role of the subbredit in the short squeeze [11]. Furthermore, other research already uses twitter data, as e.g. twitter posts with corresponding hashtags, to measure the interest of the online community concerning the GME stock [17].

In line with this research, this paper computes a number of features to determine the influence of the subreddit on the stock price of GME. Other research focuses on the general role of social media in coordinating such a mass event, and explains its counter-hegemonic motivation [4, 6]. Long et. al. [10] studies the sentiment of the comments published on r/WallStreetBets to trace its role in the development of the short squeeze. To the best of our knowledge, no research in the direction of an organic computational analysis of r/WallStreetBets in the context of metoric rise in the GME stock price has been conducted before.

In the broader context of Reddit, previous research has drawn upon methods of Social Network Analysis (SNA), sentiment analysis, topic detection and regression analysis to determine the behavior of a specific subreddit or the interconnections among reddit communities [5, 12]. Similarly, this analysis makes use of SNA and regression analysis to capture a number of characteristics of the subreddit.

Lastly, Boylston et. al. [3] analyzes the dynamics of the r/WSB subreddit. Among the dynamics considered, it reconstructs the source of trust that is established among members, evidencing the different roles inside the community and characterizing the specific jargon used.

Project Report, June 2021, Lausanne, VD, CH

Field	Description
Author	User ID of post/comment author
Created UTC	Timestamp of creation data and time
ID	Unique object identifier
Score	Post rating based on up/downvotes
Subreddit ID	Subreddit a post/comment was made in
All awardings	Lists awards given to the post/comment
URL	URL of linked content
Over 18	User/moderator tagged sensitive content
Link Flair Type	User-marked flair or topic for the post
Parent ID	ID of parent post/comment for comments
Seft Text	Text body of text post
Body	Text content of comment

Table 1: Metadata fields available in reddit dataset

3 DATA

3.1 Data Collection

To analyze the interactions within a Reddit community (subreddit) for determining their effect on stock prices, we collected publicly available data from Reddit using a third-party API (PushShift). The data corresponds to publicly visible posts and comments made by users on the subreddit. The collected data covers the period of highest volatility of the GameStop stock price, extending from January 1^{st} 2021 to March 17^{th} 2021. Financial data describing the hourly stock price of GME was collected for the same time period using the Yahoo finance API [2].

Introduction to Reddit data. Of the different interactions between users available on Reddit, posts, comments, upvotes and awards are publicly available. Other interactions, such as direct messages, are private and therefore not available via the Reddit API. Interactions happen within communities (aka subreddits) dedicated for a particular topic (for e.g., music, gaming and videos).

Users initiate activity within a community by authoring posts. Posts come in two varieties: link posts and text posts. Link posts contain a title and a link to some content, often images, videos, news stories or other posts. Text posts contain a title and text written by the author. The contents of the text are generally unrestricted, and can contain links too. Posts which link to images and videos often contain content which is humorous in nature, such as memes. A significant portion of posts in many subreddits comprise memes.

Text posts, or those linking to news articles are often more serious, and used for either sharing news which concerns the community, or sharing personal opinions, explanations and analyses.

For each post on a subreddit, user interactions continue in the form of comments which form a tree structure. Top-level comments directly address the post, whereas children are comments on previous comments. Comments are free-form text fields and can contain discussions, links or both.

Users are allowed to express their like or dislike of posts and comments with the up/downvote feature. Each user is allowed to contribute a single up/downvote to every comment or post. Posts and comments, therefore, have a score which is the resulting difference of upvotes and downvotes. Visitors to subreddits often sort posts by "Top" (highest score within a period) or "Hot" (score/time), both of which result in posts with high scores gaining higher visibility.

Finally, communities contain moderators who work to enforce community policies. As part of their arsenal, they are able to delete posts and comments which violate rules. Users are also allowed to delete their own posts and comments.

Basic data description. We collected data using the API available from pushshift.io, which mirrors reddit data. Compared to the API natively offered by Reddit, this enabled us to search for posts within a start and end-point in time. Both Reddit and Pushshift enforce a limit of 1000 items in the response for any query. Having the ability to dynamically query posts within a time-frame allowed us to download the entire post history ($900'100$ posts) over multiple queries.

The scoring system of Reddit also ensures that a vast majority of posts receive little to no attention. From the list of all posts, we choose to ignore posts which have less than a single comment and a single upvote. We believe that such posts have no impact on the community, and can be safely ignored without affecting our analyses. Of the $138'364$ posts which remain, we also discard posts which have been deleted. Unfortunately, it is entirely possible that a deleted post enjoyed a period of popularity before its deletion, and that it affected the community sentiment. However, the post's data are not available from either the Reddit or Pushshift API, so our data set remains lacking in this respect. Finally, we have $80,283$ posts remaining, used for the analysis.

For each post, we also fetch its list of comments. In total, our data set include $15,371,115$ comments. Fetching comments include separately querying for children of every previously fetched comment, and is a time-intensive process. Each post and comment object includes metadata, of which particular fields are shown in Table 1. A basic exploration of the data, and summary statistics are shown in subsection 5.1.

4 METHODS

4.1 Social Network Analysis

Social Network Analysis (SNA), as shortly covered in the Research Context, is a widely adopted technique that emerged from the social sciences and nicely lends itself to the study of online social networks [9]. SNA is based on the study of interactions between players inside a system and is enriched by a large variety of statistics designed specifically for the technique. In this study, the network is built considering the users of the subreddit as nodes and tracing an edge for each connection between one user and another when the first user has commented on the other user's post. In this sense, the paper adopts a directed graph as basic network structure.

The goal of this method is to answer the first research question. Investigating connections between users is here aimed at determining whether the structure of the subreddit creates the possibility of a mass phenomenon. To be a mass phenomenon, the subreddit needs to exhibit three necessary aspects: *i*) the structure of the r/WSB should mirror that of a tight community, *ii*) transmission of information between users should be quick, enabling trust, and *iii*) users should cover a variety of different roles, each with diverse key players.

In order to detect whether a community is tight, the following hypotheses should hold. The average degree of the network[1], encoding the average number of links from/to a user, should be high. This would indicate that users are, on average, connected to many other users and can, therefore, organize efficiently a mass phenomenon. Furthermore, the overall density of connections in the network (the number of actual connections over the total possible connections) should also be rather high, showing that the amount of total connections is suited for a large mobilization of users. The number of reciprocated links should also ideally be high as it manifests the presence of bidirectional interactions during the event. Finally, similarly to real world networks, a giant component is to be expected that demonstrates that most of the users in the network have interacted with each other [9]. The global amount of connections should also be accompanied by a high clustering coefficient (number of closed triplets over the number of triplets in the network), for which even local structures are highly connected.

For the short squeeze to have been a mass phenomenon, numerous members of the community need to have bought shares. For this to have happened, the users need to have trusted in this investment. This paper assumes that one of the reasons that pushes users to trust in the investment is if the members promoting it are, in some way, related to the users themselves. This implies that the distance from one user to another should on average (average shortest path) be low, so that a user is more likely to have come in contact with the people posting about the investment. Accordingly, the maximum distance (diameter) should also be low.

Although it is to be expected that a few users were most active in the community and in the event, to determine whether these were 'sharks' who monopolized the discourse or merely some more proactive users, this project bases its analysis on the types of key players and their roles in the network. Firstly, we assume that the discourse over this event should be created at different levels inside the community: gatekeepers should be spreading the information, visible figures should be promoting it, and common users reacting to it and 'pumping the hype'. Two metrics detecting the gatekeepers of the community were used. The first is closeness centrality, which ranks users by the average number of steps required for them to reach every other user in the network and therefore yields the users who are closest to all other users; the second is betweenness centrality, which returns users who appear most often in the shortest paths between any two users. These encode two different types of gatekeepers and should, therefore, also extract different central figures in networks that are characterized by different key players. To detect which were the visible figures in the event the HITS algorithm was used [8]. This detects the authorities of a network (users who are often linked to, and who are, therefore, visible), and hubs (users who link to many other users, representing the most active 'common users'). Comparing among the resulting sets of popular users and determining how different these are is considered by this paper a means to determine how varied the number of proactive users in the event actually was.

4.2 Time series Forecasting with Ordinary Least Squares (OLS)

To assess the cross-correlation and a possible relation between the activity on the r/WSB subreddit and the meteoric rise in the stock prise of the GameStop (GME) stock, an ordinary least squares (OLS) regression model is employed. OLS regression models are commonly used in economical and financial analysis to predict time series or more specifically to 'forecast' time series [14]. Although OLS models used to forecast time series do not offer any causal interpretation, they still show the relation between time shifted variables and might encode interesting information [14].

To analyse the effect of the activity on the r/WSB subreddit onto the price change of GME, we regress some features modeling the activity in the financial forum onto a time lagged price change. The most granular financial data at our disposal is the hourly stock price of GME. The resulting time series thus consists of the hourly change of the GME stock price. To ensure that the time series is not stationary[2] a Augmented Dickey-Fuller test has been conducted [13], which indicates that our time series is indeed non-stationary (p-value: 0.000588).

To model the time dependence between our independent variables X and our dependent variable y correctly, a time lag of one hour has been implemented. Thus our resulting regression models the effect of our explanatory variables X_t in time t onto our dependent variable y_{t+1} in time $t + 1$. Given that the stock marked is closed in the night and during the weekends, the explanatory variable X does not always encompass the same period of time. Time fixed effects, e.i. weekend and night fixed effects, are use to control for this difference in the time span of observation.

The resulting regression then measures the effect of activity on r/WSB in period t onto the price change of the GME stock in period $t + 1$. To model the activity on r/WSB in period t several features have been created. The most naive feature consists of the aggregated number of post discussing the GME stock in period t. To filter post which discuss GME, only post have been consider which clearly mention GME, GameStop or another GameStop synonym in their titles (in total 24'919 posts). In another regression the post have been differentiated according to their self flagged topic (see Table 1), in either discussion post (in total 8'110 pots) or reactionary posts[3] (in total 16'809 pots). Here we assume that posts might have a different effect according to the nature of the post. In another regression specification, the posts are weighted according to the number of comments they generate. Therefore, we weighted the number of discussion and reaction posts in period t by interacting them with the z-normalized number of comments they cause. A last regression specification also comprises two more covariates, which are the z-normalized post score and the average sentiment of the post title as determined by the NLTK Vader sentiment analyzer [7]. The final regression formulation is described below:

$$\begin{aligned} \Delta GME\,Price_{i,t+1} = \beta_0 + \beta_1 Discussion_{i,t} + \beta_2 Reaction_{i,t} + \\ \beta_3 Comments_{zNorm,i,t} * Discussion_{i,t} + \\ \beta_4 Comments_{zNorm,i,t} * Reaction_{i,t} + \beta_x X_{i,t} + e_{i,t} \end{aligned}$$

[1] For more information on the network statistics mentioned in this section, refer to [9]

[2] A stationary time-series does not have a time-dependent structure, potentially leading to a spurious regression

[3] Consisting of all other self tagged flair beside 'Discussion'

Where $\Delta GME\,Price_{i,t+1}$ is the price change in the GME stock in one hour, $Discussion_{i,t}$ is the absolute number of discussion post in period t, $Reaction_{i,t}$ is the absolute number of reaction post in period t, both variables are interacted with the z-normalized number of comments denoted by $Comments_{zNorm,i,t}$. $X_{i,t}$ signifies the aforementioned covariates and $e_{i,t}$ the error of the model.

5 RESULTS

5.1 Summary Statistics

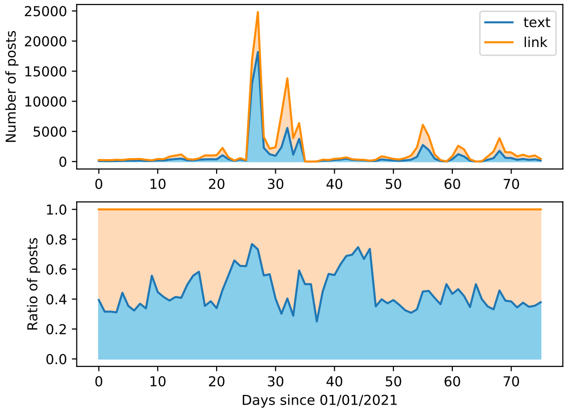

Figure 1: Number and types of posts per day.

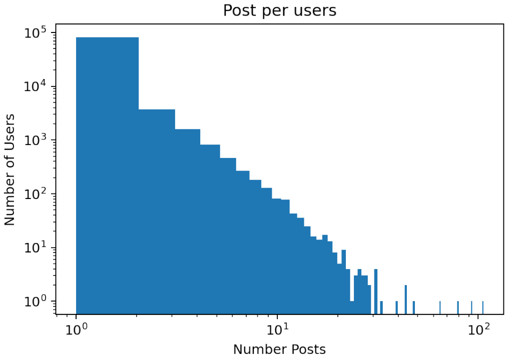

Figure 2: Complementary cumulative distribution function of posts per user.

This section briefly characterizes activity on the subreddit over the period of collected data. Figure 5 plots the number of posts per day over the observation period. We can see a massive rise in the number of posts following the short squeeze, around day 30. This period of high activity also was the period of the highest volatility of the stock's price. During this period, a high fraction of posts were text posts, indicating that the subreddit's userbase wanted to make their voices heard. However, we can see that the vast majority of users make a low number of posts, and that only a few hundred users make more than 10 posts (Figure 2). More summary statistics and plots can be found in the Appendix.

5.2 Network Analysis

Following the specifics in the Methodology, a directed network with 637'395 users and 2'039'746 links (comments on posts) was built. This included on average 3.20 comments per post. Due to computational limitations, very few statistics could be run on the full network in a reasonable time-span. For this reason, a subset of the network was selected. A section of this subgraph can be seen in Figure 3. This was done sampling 30'509 random users and inducing the subgraph on these nodes. The resulting network has 103'761 links and 3.40 comments per post. This method was chosen as it maintained the computed statistics (such as density, average degree, average shortest path) similar to those of the original network.

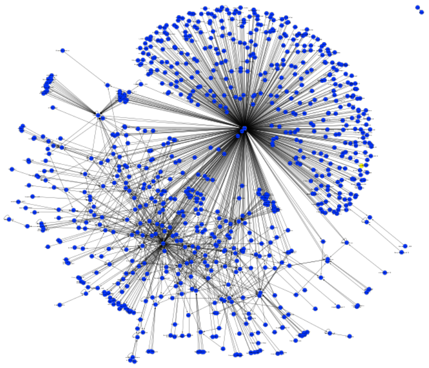

Figure 3: Visualization of a random subset of the network.

On this subnetwork, the three components which characterize for us a mass phenomenon were analyzed. All the statistics of the network were computed using the python library `networkx` and reported in Table 5. Regarding the first set of hypotheses, the average degree computed is 3.40. Looking at Figure 4, the degree distribution of the network follows a Zipf distribution (here note that the plot uses both logarithmic axes). This indicates that, while few users (possibly bots and moderators) have extremely high degrees, most users have about 0-10 neighbors. Being the average number of connections between 3 and 4 users and considering that this stays largely above 1 even with all the novices in the network who have none to 1 connections, this number appears rather high. However, the density of the network is barely 0.01 and the reciprocity rate 0.01. These two measures, although contradicting the hypotheses of a tight community, can be explained by the nature of the subreddit, where users do not know each other on a personal level and are, therefore, unlikely to interact with each other if not for interest in a specific post. Tracing the connected components of the network, a giant component of 23'870 nodes[4] was detected, in accordance to the hypothesis. Moreover, the clustering coefficient is 0.23, indicating that, at a local level, about 23% of the connections between any three users of which two are connected are complete. This is rather high considering that, in a random graph, this would be barely $\frac{3.4}{2039746}$ [9].

[4] This is to be expected a the average degree is greater than 1 and lower than $ln(N)$ where N is the number of nodes

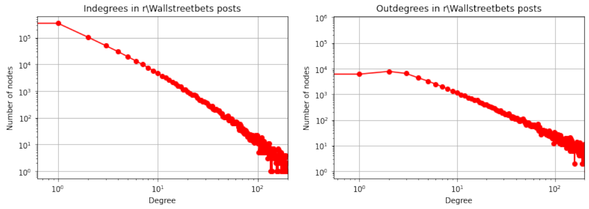

Figure 4: Left: In-degree distribution on a loglog plot, Right: Out-degree distribution on a loglog plot.

	Full-graph random subgraph
Number of nodes	30509
Number of edges	103761
Average degree	3.40
Diameter	8
Average clustering	0.23
Reciprocity	0.01
Average shortest path length	2.67
Number of communities	6626
Mean size of communities	4.6

Table 2: Table containing statistics for the subnetwork.

The two metrics that encode the speed of the 'trust' channel in the network are both largely in accordance to the hypotheses. The maximum separation between two users in the subreddit is only 8 steps and the average is only 2.67. This is remarkably small when compared to Milgram's coefficient [16], 6, and to the coefficient in Online Social Networks where it is about 4 [15].

Finally, the results on the roles and key players of the network were the most arduous to interpret. Table 3 shows the most central nodes according to the four centrality measures listed in the Methodology. The two measures encoding gatekeepers identify somewhat different users. Furthermore, the visible figures seem to be a summary of the most important gatekeepers. These include, among the top 20, Keith Gill (deepf******value), who was certainly one of the key actors in the network. These central users, when looking at actual posts, tend to orient the discussion towards recapitulations of the market, comment on important headlines and communication of crucial information to the community. The common users identified by the hubs are completely different from the others and, looking closely at the types of posts by these users, they reflect on their emotions, they show hype for GME and talk about their investments. The fact that the key players are different for each metric and cover different roles to propagate information, is a good indication that the community is not ruled just by few 'sharks', but rather different actors participate in different ways in the event. Looking more in depth into the community, one can see that moderators and bots cover very prominent roles in the community (as can be seen in Table 3, where the first two positions are always covered by either of the two). These two roles, interestingly, appear in all four metrics, indicating that the two roles are not most central in either specific way, rather, they are as present inside the network in different ways just as normal actors of the network.

Degree centrality	Closeness centrality	Authorities	Hubs
MotorizedD*****Canoe	MotorizedD*****Canoe	MotorizedD****Canoe	*MangoManYummy*
theycallmeryan	**zjz**	*theycallmeryan*	*nosalute*
Bundaga	Tradergurue	**zjz**	giantwashcapsfan8
speakinexistencebro	Tradergurue-Prize	junaidminshad	FaithlessnessFree331
Tradergurue	MIA4real	MIA4real	disneysinger

Table 3: Table containing top central nodes of the network. Not considering OPINION_IS_UNPOPULAR and AutoModerator which are always the most central (except for Hubs). The first is a moderator while the second is an automatic moderator. When a user is a moderator in the table it is bolded, when it is a bot it is highlighted in italics.

5.3 Time Series Forecasting Results

By implementing the described time series forecasting OLS regression model we obtain Table 4. Table 4 describes four different estimated regressions with an increasing number of features used to explain the dependent variable. The first regression indicates that there is a significant positive effect of the number of GME posts in period t onto the price change in period $t+1$. However, this effect seems to be rather small. The second regression differentiates the type of post into discussion posts and reaction posts. Here it becomes evident that discussion posts have a significant positive effect, while reaction posts have a significant negative effect onto the price change in the next hour. Interacting the discussion and reaction posts with the z-normalized number of comments per post in the third regression increases the magnitudes of the discussion and the reaction posts coefficients. Thus we see that per discussion post with one standard deviation more than average comments the stock price increases by 2.7 dollars. The last regression uses the in the method section explained covariates to explain the change in stock price. In this regression all discussed coefficients point into the same direction and are still highly significant, they decreased however in magnitude.

Analyzing the obtained results, we observe some interesting trends. First, we clearly see that post on the r/WSB subreddit mentioning GME have a significant effect onto the change in stock price in the next hour. This effect seems to differentiate according to the type of the post. Discussions seem to lead to an increase in stock price while reactionary post are correlated with a decrease in stock price. Furthermore, the magnitude of those effects seems to strongly vary with how much attention the posts get, as expressed by the amount of comments on a given post. While additional covariates explain additional variance, they are not significant.

Although the produced results do not claim any causality, they still indicate some interesting relations. The produced results suggest that serious financial discussions on the r/WSB subreddit might have motivated people to invest into the GME stock. Reactionary posts however seem to be correlated with a decrease in stock price, which might be explainable by the fact that people tend to post after a significant price increase, which is just the moment when the price decreases again. Lastly, the magnitude of those effects seems to partly be explainable by the attention such post receive, as measured by the number of comments on a given post.

	Dependent Variable: Hourly Change in Stock Price GME (N=350)							
	GME Mentions	p-value	Differentiation of Post type	p-value	Differentiation of Post importance	p-value	Differentiation of Post importance & Covariates	p-value
GME Posts	0.0092***	(0.003)						
GME Discussion Posts			0.1686**	(0.027)	0.9120***	(0.000)	0.6975***	(0.000)
GME Reaction Posts			-0.0830**	(0.016)	-0.4741***	(0.000)	-0.3610***	(0.002)
Comments z-Normalized					-0.0890	(0.918)	-0.0621	(0.943)
Comments z-Normalized $\ast GME\ Discussion\ Posts$					2.7308***	(0.000)	2.0279***	(0.003)
Comments z-Normalized $\ast GME\ Reaction\ Posts$					-1.4197***	(0.000)	-1.0802***	(0.005)
Weekend/Night Fixed Effects	✓		✓		✓		✓	
Other Covariates	✗		✗		✗		✓	
Df Model	3		4		7		13	
R-squared	0.057		0.147		0.203		0.230	
Adj. R-squared	0.049		0.137		0.186		0.200	
F-statistic	7.020		14.81		12.41		7.702	

Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 4: Results of OLS Regression: Forecasting activity on r/WSB onto GME stock price change

6 DISCUSSION

The results obtained by the network analysis indicate that r/wsb is a tightly knitted community with short and fast communication channels and many different key players. Thus we believe that the r/wsb community is not dominated by a few influential actors, but rather characterized by many diverse participants. This would indicate that the trend concerning GME was a mass phenomenon. Assessing the results from the regression analysis, we find that post discussing GME seem to have a strong correlation with a price increase in the GME stock. This effect is amplified if a given post is well discussed and receives a lot of attention from the community as measured by the number of comments.

Although our analysis yields convincing results, this paper suffers from several limitations. First, although we analyze more then 80'000 distinct post, this is only a fraction of the 900'000 total post which were posted to r/wsb in the discussed time period. However, to comply with storage and computational restrictions, all posts were excluded which had less than one upvote or comment or were deleted, which leads to the resulting data set of 80'000 posts. Furthermore, this analysis only considers a time frame from January 2021 to mid March 2021, which might not reflect the phenomenon in its entirety.

The network analysis also comprises limitations. Most importantly, it could not be run on the full network. Additionally, the metrics computed did not consider edge weights, nor comments on comments. Missing edge weights imply that the number of comments between two actors does not influence their connection, but an edge indicates only if one user ever commented on another user. Furthermore, comments on comments are not considered in our analysis. The analysis was run on a random subset of the full graph which might influence or even bias the obtained results. Finally, to ensure interpretability, more precise comparisons to random graphs should be made on the measures (such as degree and clustering coefficient) to be able to more accurately determine whether these are high or low.

The main flaw of the regression analysis is that it is purely correlational and cannot claim any causality. The project could be extended by introducing methods which ensure causality (e.g.: instrumental variable regression). Furthermore, given that all regression features are self constructed, it might also be that they are ill constructed and other variables are better suited to explain the variance in the data. Lastly, given the high influence of the number of comments on the magnitude of the coefficients, an interesting future step is to further investigate the content and nature of those comments.

7 CONCLUSION

In conclusion, the network analysis has shown that the community is suited for a mass phenomenon. This is due to its tight connections, the closeness between users which elicits trust and the participation in the community in many different way. The results of the discussed time series regression analysis indicate that there is a strong and significant correlation between activity on the subreddit r/WallStreetBets and changes in the stock price of GameStop. Although this regression analysis claims no causality, the produced results still indicate a relationship determined by complex functionalities between the online financial forum and the real stock price of the GameStop cooperation.

REFERENCES

[1] [n.d.]. GameStop Short-Sellers Reload Bets After $6 Billion Loss. https://www.bloomberg.com/news/articles/2021-01-25/gamestop-short-sellers-reload-bearish-bets-after-6-billion-loss

[2] David Berube. 2007. Retrieving Stock Quotes with YahooFinance. *Practical Ruby Gems* (2007), 103–107.

[3] Christian Boylston, Beatriz Palacios, Plamen Tassev, and Amy Bruckman. 2021. WallStreetBets: Positions or Ban. *arXiv preprint arXiv:2101.12110* (2021).

[4] Riley Burnette. 2021. What Were the Factors that led to the GameStop Short Squeeze? (2021).

[5] Bryce Cai, Sean Decker, and Crystal Zheng. [n.d.]. Characterizing Banned Subreddits by Network Attributes. ([n. d.]).

[6] Usman W Chohan. 2021. Counter-Hegemonic Finance: The Gamestop Short Squeeze. *Available at SSRN* (2021).

[7] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8.

[8] Jon M Kleinberg. 1999. Hubs, authorities, and communities. *ACM computing surveys (CSUR)* 31, 4es (1999), 5–es.

[9] David Knoke and Song Yang. 2019. *Social network analysis*. Sage Publications.

[10] Cheng Long, Brian M Lucey, and Larisa Yarovaya. 2021. " I Just Like the Stock" versus" Fear and Loathing on Main Street": The Role of Reddit Sentiment in the GameStop Short Squeeze. *SSRN Electronic Journal* (2021), 31.

[11] Štefan Lyócsa, Eduard Baumöhl, and Tomáš Vŷrost. 2021. YOLO trading: Riding with the herd during the GameStop episode. (2021).

[12] Markus Moessner, Johannes Feldhege, Markus Wolf, and Stephanie Bauer. 2018. Analyzing big data in social media: Text and network analyses of an eating disorder forum. *International Journal of Eating Disorders* 51, 7 (2018), 656–667.

[13] Rizwan Mushtaq. 2011. Augmented dickey fuller test. (2011).

[14] James H Stock and Mark W Watson. 2015. *Introduction to econometrics*.

[15] János Szüle, Daniel Kondor, Laszlo Dobos, Istvan Csabai, and Gabor Vattay. 2014. Lost in the City: Revisiting Milgram's Experiment in the Age of Social Networks. *PloS one* 9, 11 (2014), e111973.

[16] Jeffrey Travers and Stanley Milgram. 2011. An experimental study of the small world problem. In *The structure and dynamics of networks*. Princeton University Press, 130–148.

[17] Zaghum Umar, Mariya Gubareva, Imran Yousaf, and Shoaib Ali. 2021. A tale of company fundamentals vs sentiment driven pricing: The case of GameStop. *Journal of Behavioral and Experimental Finance* 30 (2021), 100501.

[18] Evangelos Vasileiou, Eleftheria Bartzou, and Polydoros Tzanakis. 2021. Explaining Gamestop Short Squeeze using ntraday Data and Google Searches. (02 2021).

ADDITIONAL FIGURES

Flair	Number posts
Discussion	43’494
YOLO	28’179
News	18’008
Meme	13’081
Gain	11’807
Loss	9’233
Due Diligence	6’210
Chart	4’171
No Tag	3’160
Technical Analysis	578

Table 5: Ten top self labeled topics or ’flairs’ with the most posts

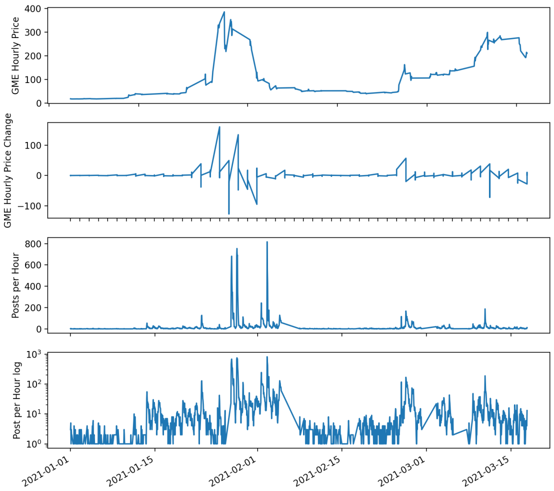

Figure 5: Time series of a) GME stock price hourly b) GME stock price hourly change c) number of post per time period d) log number of post per time period